EDA-Ames Housing DataSet Analysis and Linear modelling.

Ambika Huluse Kapanaiah(u3227622)

09/05/2021

1. Title and abstract:

The Ames Housing Data Set is popular among Data Science world. Everyone is challenged to work on minimizing the RMSE value on test set. Similarly,in this report we study the dataset by applying various Exploratory Data Analysis and create appealing linear model with train and test data-set, which includes 81 features describing a wide range of characteristics of 1,460 homes in Ames, which were sold between 2006 and 2010. Along with this, this report also includes 10 problems which are identified as part of prediction of SalePrice based on various explanatory features. So, here we identify key features which are really affecting SalePrice, based on the information collected from various websites. All the websites referred for this project has been documented at the end of this report.

2.Information on the dataset:

Loading and understanding the data.

```
Ames_train <- read.csv("data/train.csv",na.strings=c("","  ","NA"))
Ames_test <- read.csv("data/test.csv",na.strings=c("","  ","NA"))
```

The Ames Housing data set consists of 80 variables considering SalePrice as our interest as target feature. All the 79 explanatory variables focus mainly on the quantity and quality of many physical attributes of the property.Most of the features are mainly the type that naturally any customer who would want to buy a property (For Ex: When it is built, How big the property is? How about the living area sqft? Car parking? ? How many bathrooms and bedrooms available? Materials used for flooring roof and finishing? Location of the property? and so on).Few continuous variables which were showing the data on the various dimension of the property. LotArea, PoolArea,GarageArea and so on. Few categorical variables describing the quality and type of the overall amenities and materials used to build/renovate the property/street or neighborhood and nearby amenities and so on. Few discrete variables showed the number and location of amenities/bedrooms/bathrooms/kitchens located within the property.Also few temporal variables which says about the year of renovation/Garage built/Year built for the property.After careful observation there are major features present that effect SalePrice and could be summarized as property size, number of rooms,location, amenities, constructed materials, property's overall age and condition of the property/amenities.

3. Problem Identification:

These are the few DataScience problems, we came across while studying dataset and affect SalePrice based on the key features. We will find solution to each one of them once we finish data Analysis based on several questions which we think of.

1. Problem 1: Identify which suburb/location had the biggest growth in SalePrice by plotting and examining the sale prices cross different suburbs.Has there been a trend on the type of house bought and had big hike in Sale price from 2006 to 2010

2. Problem 2: Analyze a possible pattern of SalePrice vs YrSold/MoSold, LotArea and/or some other variables which can reasonably be included considering Totl_Area instead of LotArea,SeasonSold instead of MonthSold here.

3. Problem 3: Whether SaleCondition has any impact on the SalePrice, Explain with Data Analysis and give insights on whether this feature needs to be considered.

4. Problem 4 : Any change in SalePrice over the period from 2006 to 2010 based on GarageQual

5. Problem 5:Over the years.How the SalePrice changed based on Neighborhood and BldgType .Explain

6. Problem 6:Was there any difference in SalePrice for the properties sold between 2006 to 2010 based on LotShape/LandContour? which is basically considered in Indian tradition, Just for curiosity have included this problem analysis.

7. Problem 7:Check the seasonality of SalePrice based on MasVnrType used.

8. Problem 8:Did the Age of Garage, property made any difference in the Saleprice from 2006 2010.

9. Problem 9:Do you think we could get good linear model with just considering Overall quality as a stand alone parameter for Sale Price prediction? Give thoughts.

10. Problem 10: Use predictions from your final model to compare suburbs which have shown varying growth. Or, to identify which suburbs have been growing the most over the last few years.

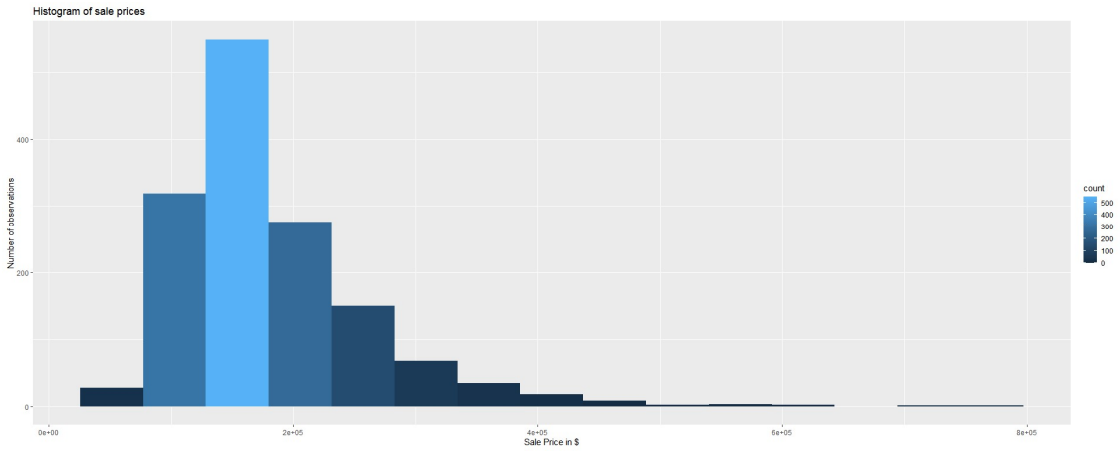
4. Project at Brief:

Basically, it looks very easy to fit a model which could say the training data is having zero error. However, that kind of model would be very poor , because the relationship between the target variable sales price and the explanatory features are not properly defined by the model. The main focus here is to fit a linear model with the best R-squared value and low RMSE. Here is the cyclic process wherein started with data exploration followed by exploratory data analysis and missing value detection and imputation. Basically, during data exploration process, one need to thoroughly study the data and relationship to the target feature. Once finding linear relationship, next step is to check the correlation values

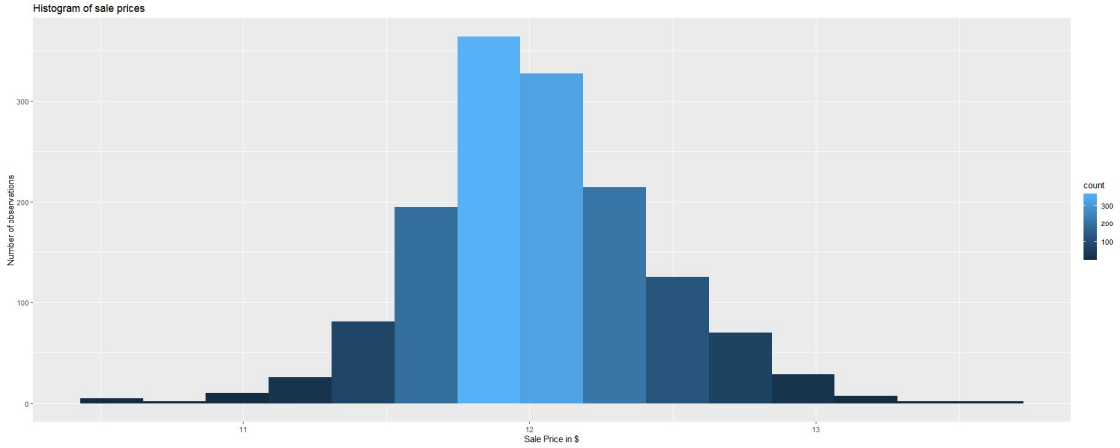
of numeric features. Here this stage answers to the question , which variables are most strongly correlated with the response. These set of numeric features will be the strongest predictor of the SalePrice. Correlation heatmap is used here to capture correlation values. Usually, deep colors on the heatmap shows strong correlation. Missing values directly affect the linear modeling. The larger the number of missing values, the poorer the model will be. If the percentage of missing value is high, that says the feature is almost not present in the property. The plots also provided to show the missing value proportion of the features. For example, the house with Pool and without Pool definitely accounts for the missing values and thereby will have discrepancies in the correlation to the SalePrice. These NA's are labeled as "Missing" in categorical features, and imputes with median value for numerical features. During feature engineering step, of Saleprice by taking logarithmic transformation done for model training. Otherwise, SalePrice was rightly skewed with few outiliers. Some features was as strings , but were ordinal features. These were were converted to numerical levels which could be further fed into models. Next, with temporal features, able to impute all the year based columns except YearSold, just to represent the age of the garage and property. This helped in showing linearity to SalePrice. Also, created SeasonSold column so as to check on the SalePrice relation with the MonthSold. Along with this, a new column with total basement area+Ground living area is created and with this new feature, able to find linearity with SalePrice again. The 10 problems which are defined above are also solved later on. Linear modeling with the best features selected by deeper study into the data set using Exploratory data analysis helped further to choose and find the best model. With this knowledge, which is gained during data analysis helped in feature selection and able to explore on 4 models with different but very near high R-squared value around 0.8, I selected features for modeling mainly based on the high correlation values and good variability of explanatory variables.And finally trained the model and fitted the curve. Using test data set the prediction is analysed and found that the chosen model is the best one among other 3 which were with lower R-squared value.

5. Data Analysis:

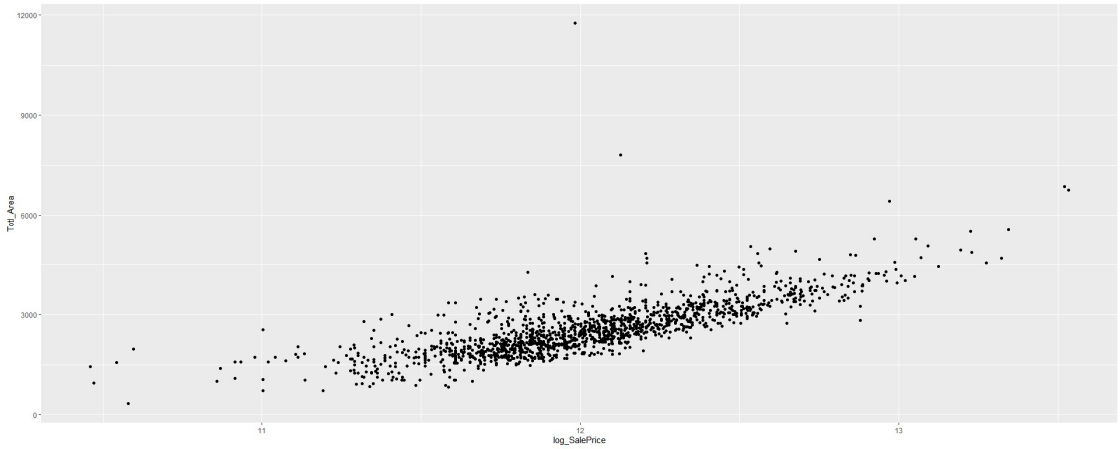
1. Every data analysis starts with checking on the distrubution statics of the target feature. The distribution of SalePrice looks skewed positive, and have outliers. We could consider log transformation in such a case.



The log transformation of SalePrice looks like normally distributed now.

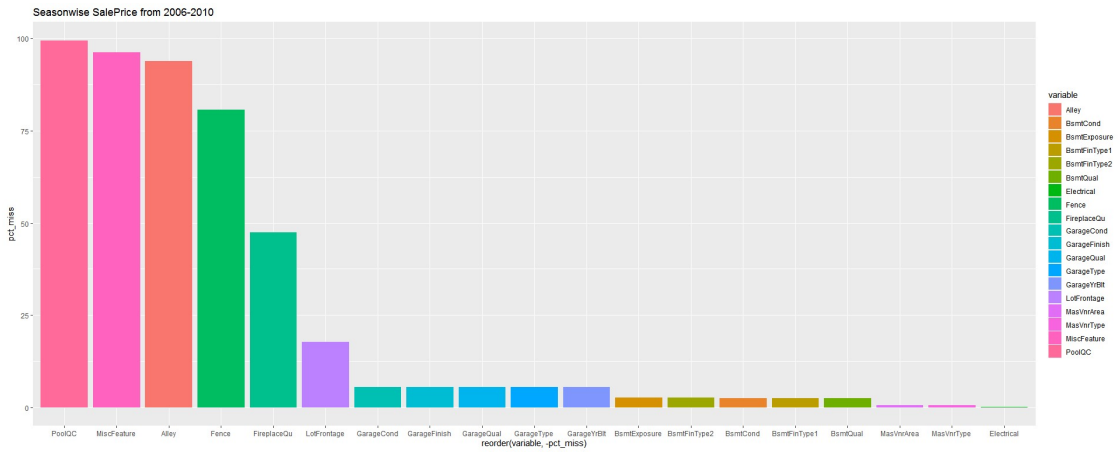


2. After careful observation, found that Total Area of the property can be added as new feature by adding features TotalBsmtSF and GrLivArea,



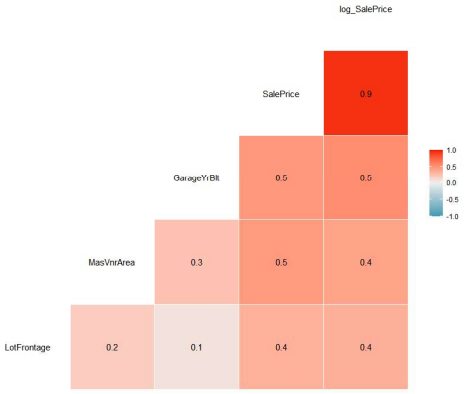
There is medium to strong linear positive relation between Totl_Area and log_SalePrice.

3. Here is the missing value representation of numerical columns.

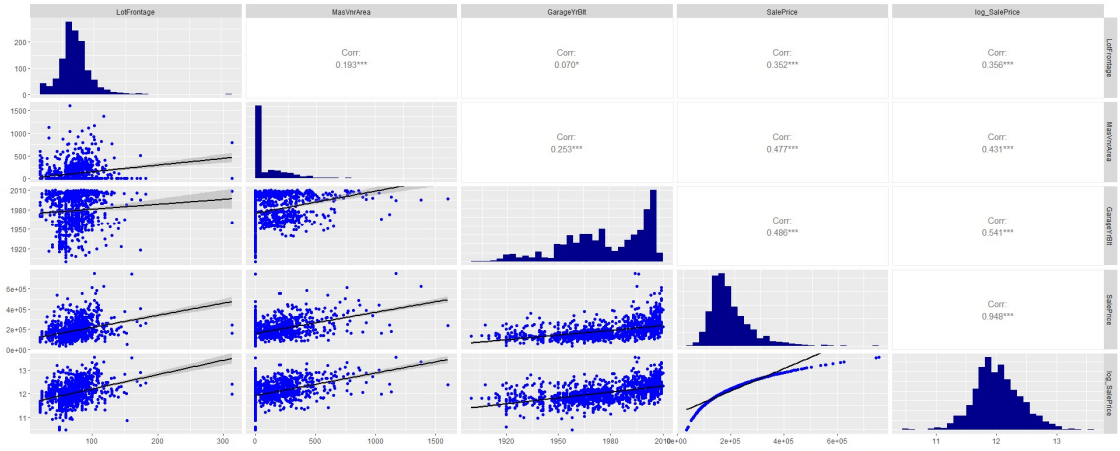


For the numeric columns with missing values, comparing correlation value here with target feature.

Considering to remove LotFrontage after careful observation. MasVnrArea and GarageYrBlt having good correlation with Saleprice.




```
FALSE `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
FALSE `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
FALSE `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
FALSE `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
FALSE `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



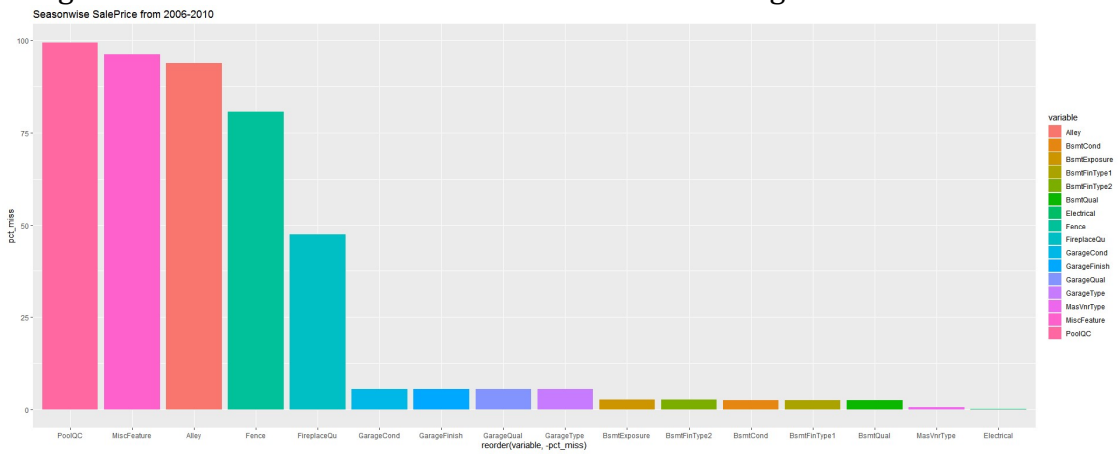
LotFrontage has less correlation , moderate to weak positive relationship to SalePrice and has got many outliers.Missing value imputation for MasVnrArea is also done in this stage. As GarageYrBlt related to Time feature, we could consider with other temporal features to check further.

Here we complete missing values imputation for Numeric columns with NA's.

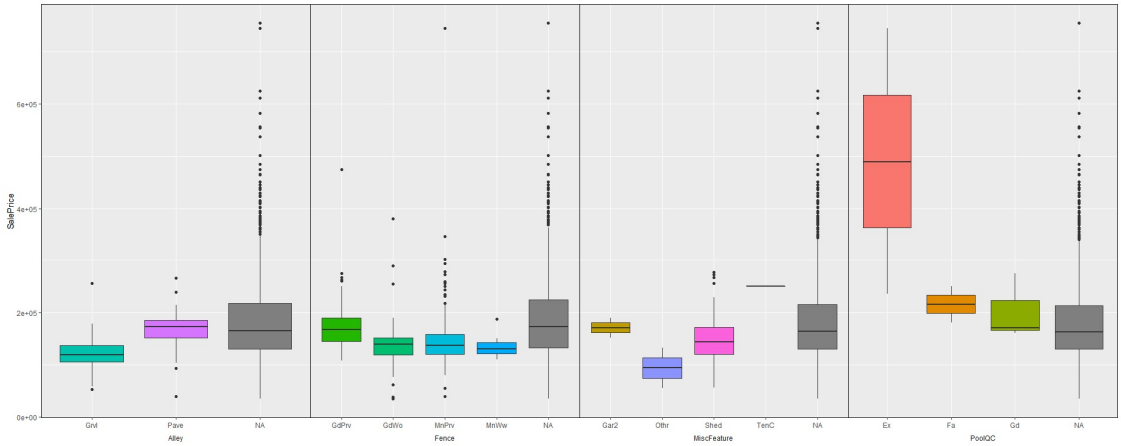
4. Here comes the bigger categoric feature with missing values

Considering only categorical columns which have missing values.

Barplot of missing percentage for each of the categoric feature is shown here. There are 3 categorical features which shows more than 80% missing value

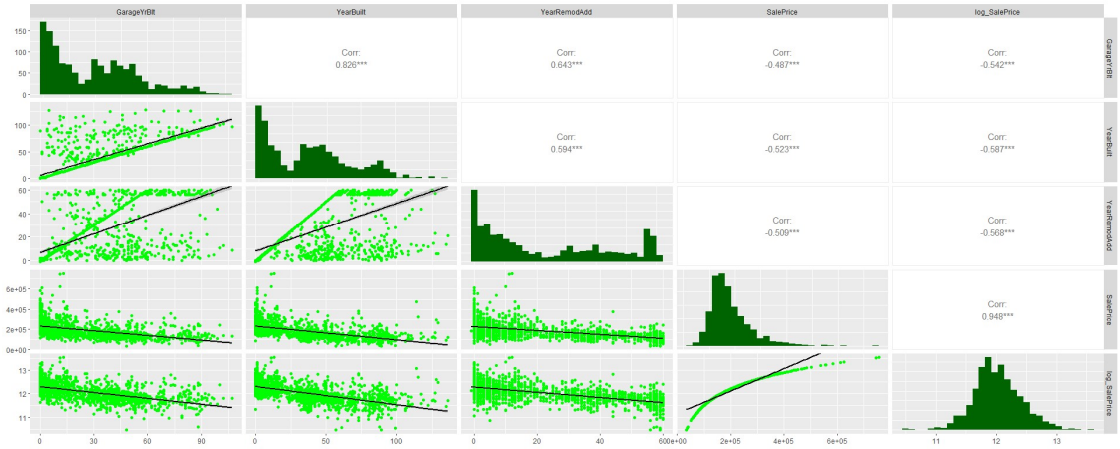



```
FALSE [1] "Alley"       "PoolQC"      "Fence"       "MiscFeature"
```



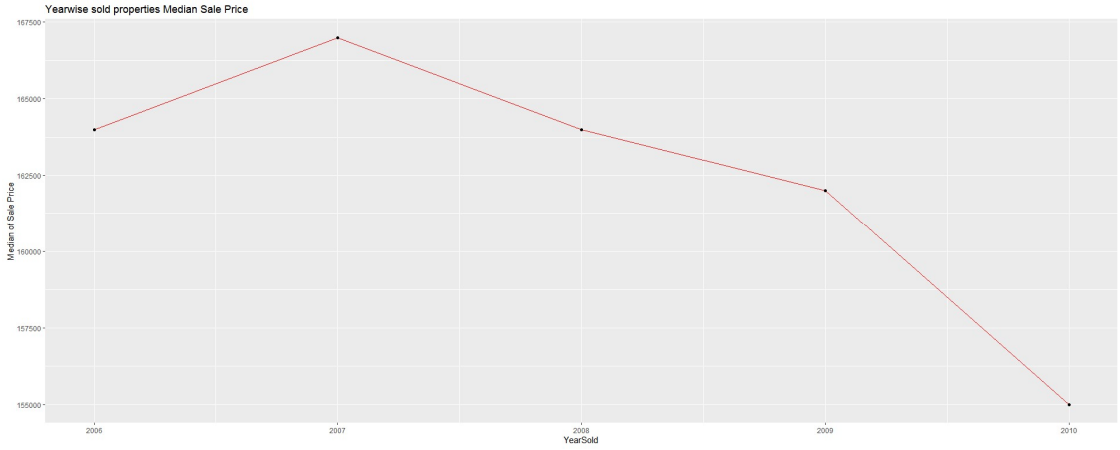
Alley,PoolQC,Fence ,MiscFeature and FireplaceQu having equal or more than 50% NA values and are having outliers. And also no much variability when related with SalePrice. We can directly drop these variables.

5. collecting all Temporal(Time/Year) related features for analysis here.After careful observation of all temoral variable, decided to convert GarageYrBuilt,Yrbuilt,YearRemod so that they represent the age of the garage and the property. As the YearSold already showing for which and all year we have data, taking difference between YrSold respectively with each feature here gives the age of the property and Garage area.



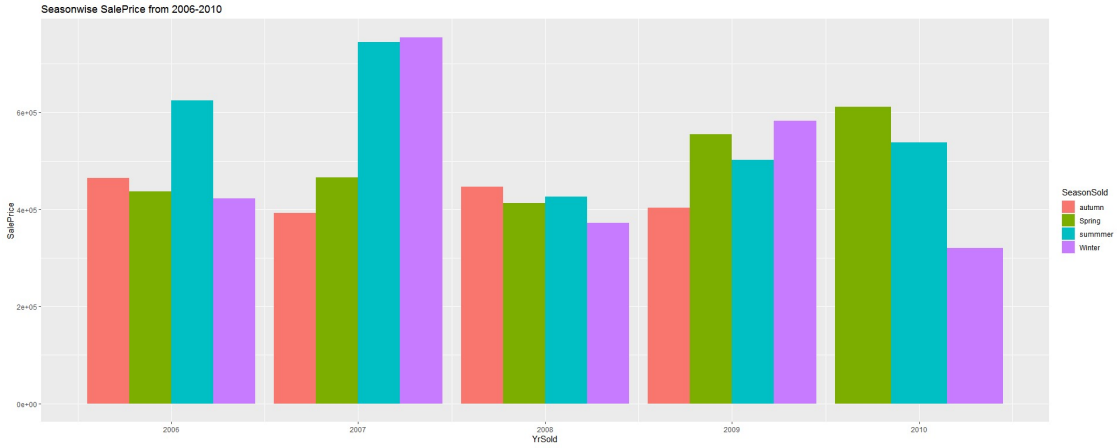
As the property ages interms of Garage or the building itself the SalePrice is decreasing. Looks like all the 3 Yearwise columns are having negative strong relationship with SalePrice and so is important in predicting SalePrice,as all their correlation value is very high.

Here checking for YrSold and SalePrice relationship.



Sale Price decreasing from 2006 to 2010, which is also an important observation to consider. Usually property prices would hike due to neighborhood and amenities. But as the age of the property increases it also takes a toss on the quality of the overall interiors and exteriors of the building.

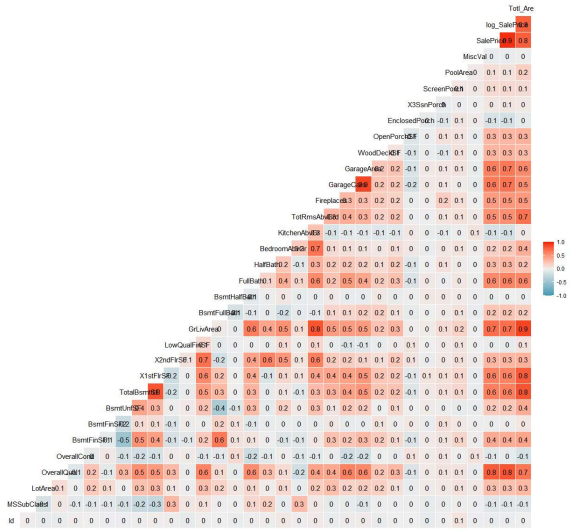
There is another interesting column which we could create here is SeasonSold. There must be a possibility based on weather, may be there is dependency to SalePrice. MonthSold could be dropped as we create this SeasonSold column here.



SeasonSold has very big impact on the SalePrice especially in the year 2007 wherein during Winter and Sumar had peak Sale Prices for the properties sold followed by Spring and Autumn.Overall all the seasons with properties sold from 2006 to 2010 are having alost common trend except 2007. Imputation of Temporal variable. Filling missing values with median is done later.

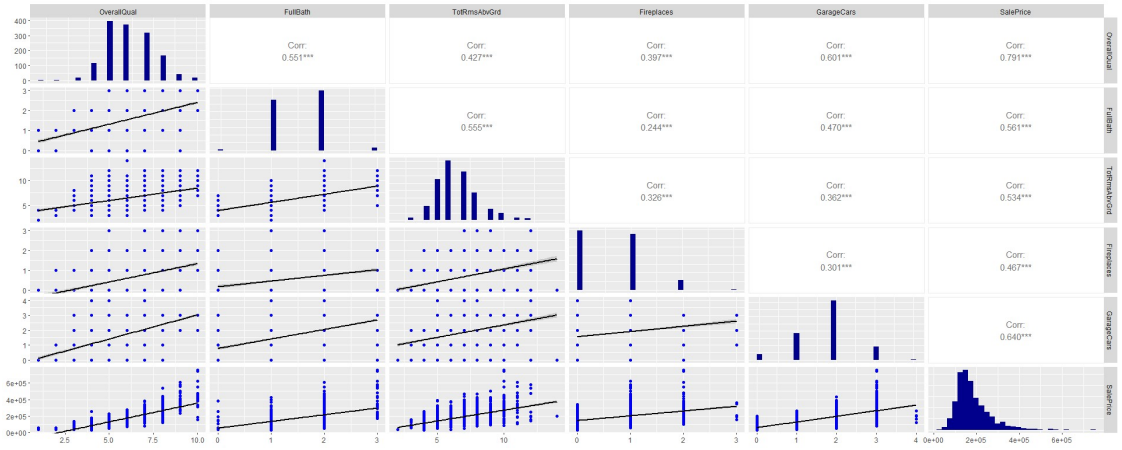
Till now explored with whether any new features can be created. And also dealt with missing values. Imputation and feature selection for features with missing values. Keeping in mind that still categorical columns are not imputed until now.

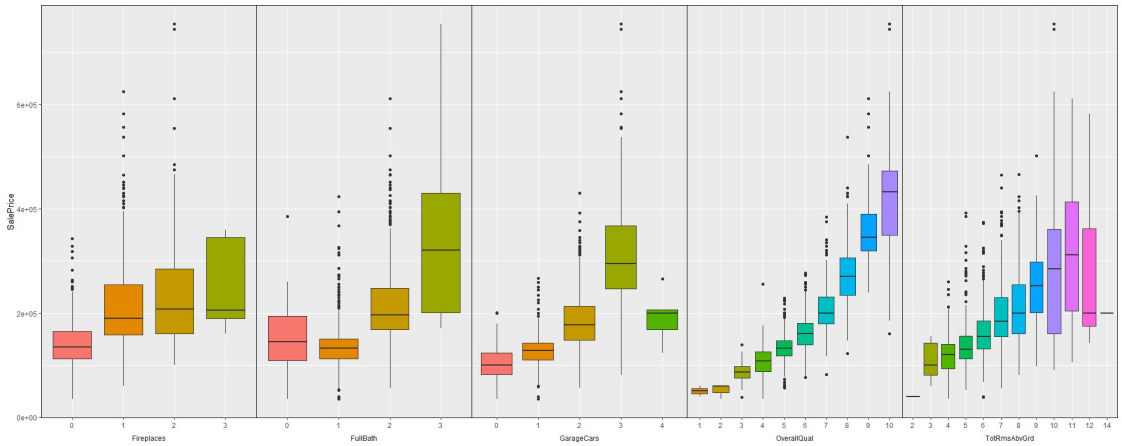
5. Here analyzing numerical columns without missing values. Whether any feature could be dropped or not is decided based on correlation values with SalePrice.



Most of the features which are having low correlation that is less than +/-0.5 with SalePrice could be dropped here.

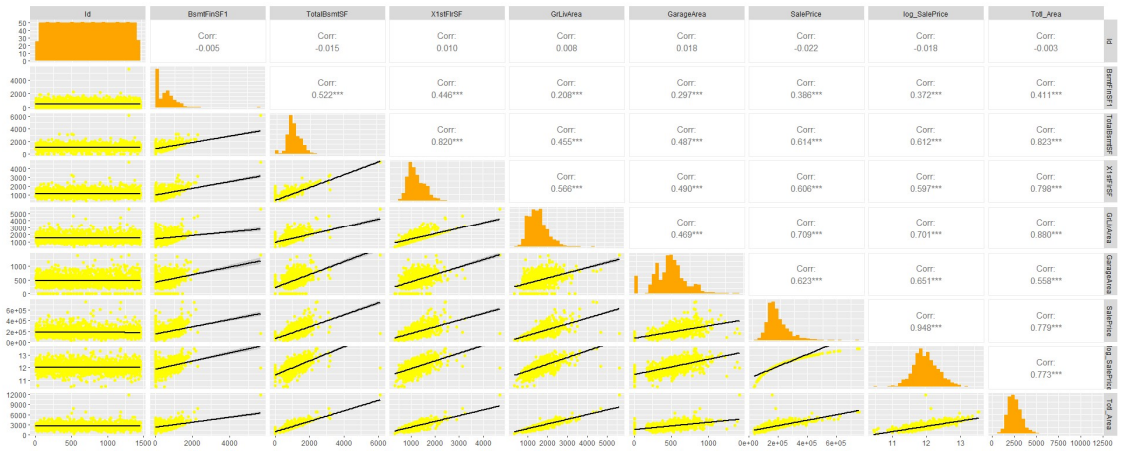
As we could see the features having less than 0.5 correlation with SalePrice shows clear info on the scatterplot which is not printed here.(Please check R code) When Analysed further segregating them to discrete and continuous features, got more insights on them.





As the overall quality,Full bath and Garage Cars having positive strong relation with SalePrice. SalePrice is increasing positively with feature Fireplaces and TotRmsAbvGrd also have positive strong relation and much variability with SalePrice

Now dealing with continuous variables with exploratory data analysis.



All these continuous features has very strong positive linear relation with SalePrice. All the numerical columns analysis completes here.

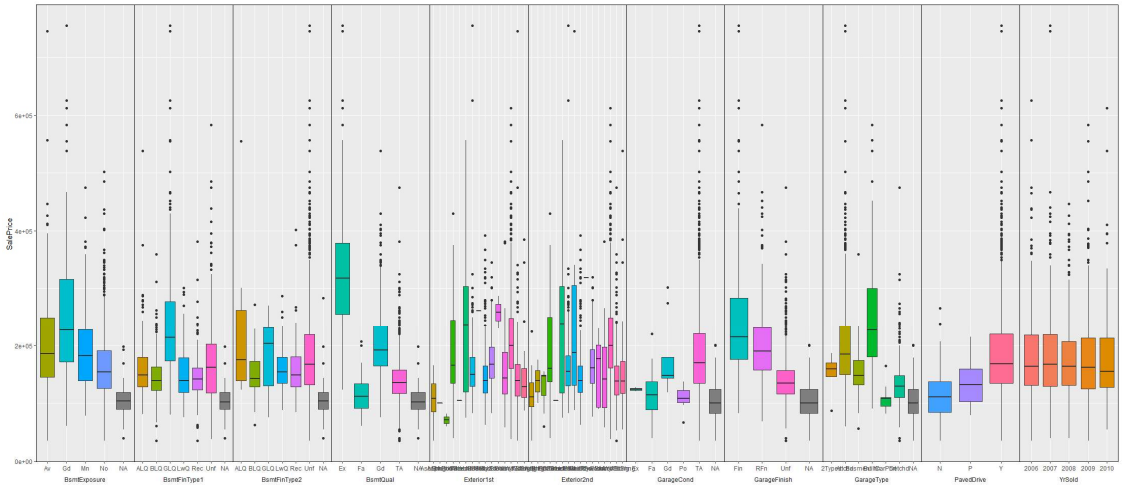
6. Now considering all the categorical variables which are not having missing values.And even those with missing values are imputed here in this step.

As there are numerous categorical features, segregated here as important categories, internal and external, miscelleneous categories. This is just for visualisation purpose.

Utilities could be removed. because of no variability with SalePrice and many outliers when checked with boxplot which is not printed here.

Four columns here can be removed as there is no much variability when compared with SalePrice: BsmtCond,ExterCond, RoofStyle,RoofMatl are dropped at this stage.when checked with boxplot which is not printed here.

TotRmsAbvGrd has very good variability compared to other features here. dropping everything except TotRmsAbvGrd here.when checked with boxplot which is not printed here.

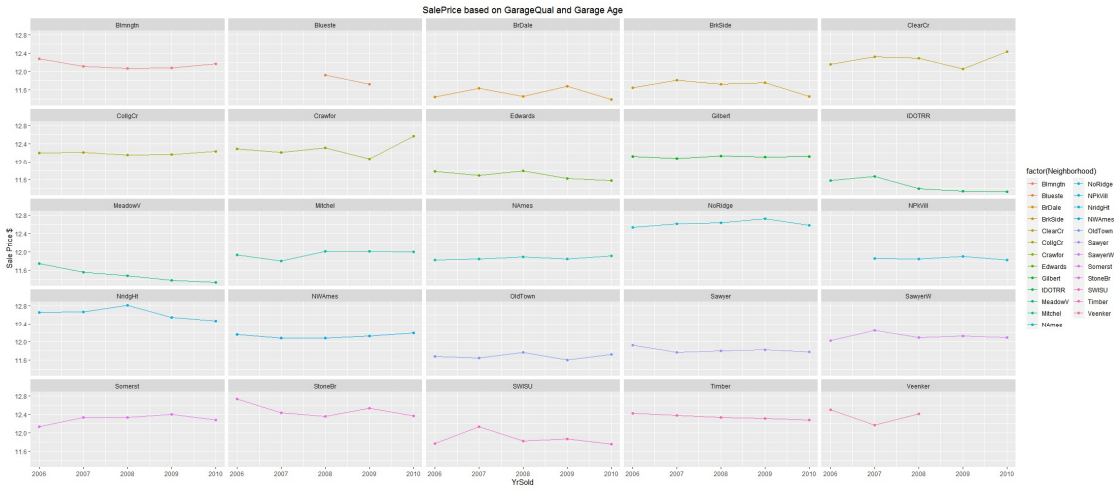


BsmtQual looks better related compared to all other features here. Removing all other except BsmtQual here. Once we select features from categorical features, Imputation for all the categorical features done at this stage.

7. Further Exploratory data analysis done at this step. Deeper understanding and finding some of the solutions to the problems identified at the beginning of this project.

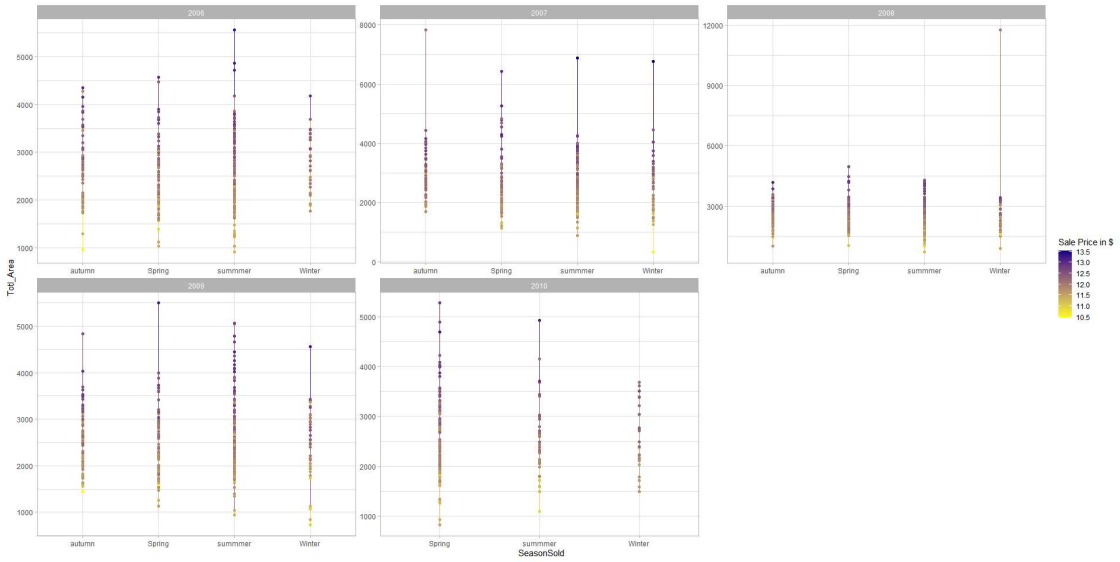
During the first stage of data analysis , the decissions are taken carefuly. so that it doesnt take much of the effort here.

Problem 1: Identify which suburb/location had the biggest growth in SalePrice by plotting and examining the sale prices cross different suburbs.Has there been a trend on the type of house bought and had big hike in Sale price from 2006 to 2010.



1) The suburb Somerst had increaing trend till 2009, in 2010 the price dropped
2) Noridge also has increasing Saleprice as trend but in 2009 it is decreased.
3) NAmes has slight increasing trend every year. So there must be other features which are affecting SalePrice here along with Neighborhood(suburb/location) of the property.

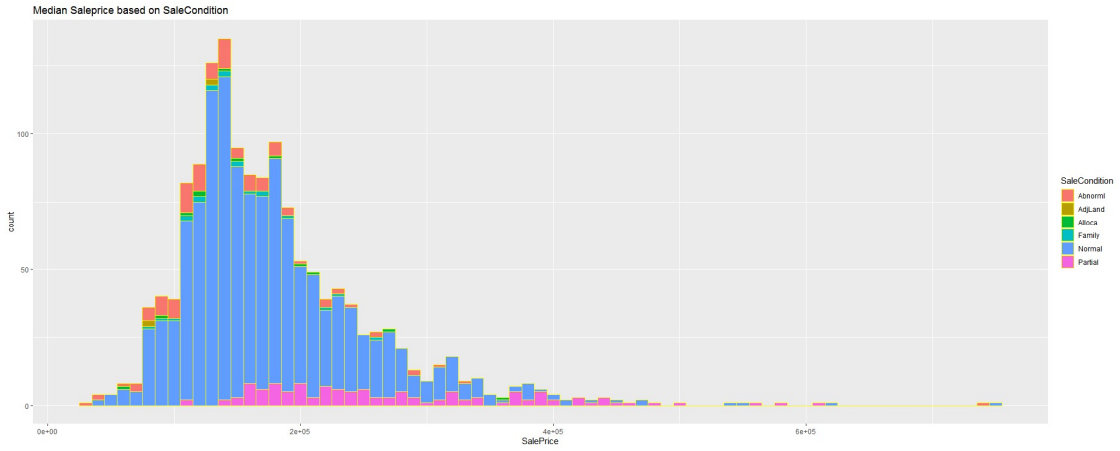
Problem 2: Analyze a possible pattern of SalePrice vs YrSold/MoSold, LotArea and/or some other variables which can reasonably be included considering Totl_Area instead of LotArea, SeasonSold instead of MonthSold

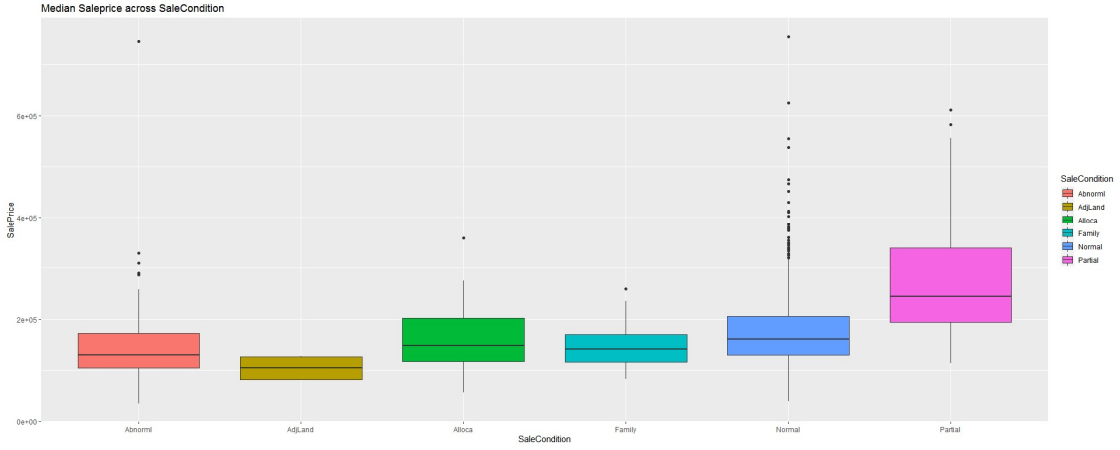


1) In 2006 during Summer the SalePrice have seen extreme peak looks like outliers, However, all the season has touched max Saleprice around 4000$ for the properties sold which are having bigger Totl_Area(Ground Living and Total Basement Area) 2) In 2008 All the

Seasons had very less Saleprices even for the properties which are big in Totl_Area, The average log_Saleprice was around 4000$ 3) In 2009 all the seasons had peaks of almost 5000$ for the properties with lasrger Totl_Area

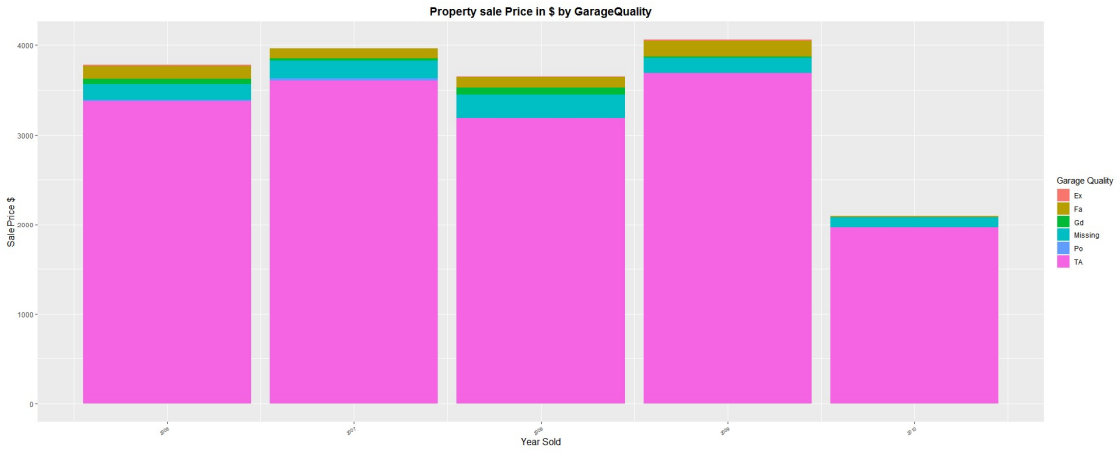
Problem 3: Whether SaleCondition has any impact on the SalePrice,Explain with Data Analysis and give insights on whether this feature needs to be considered.





1) Definitely there is a relation between SalePrice and SaleCondition
2) Normal Sale Varied with lot of outliers.
3) Family Sale between family members had extreme decrease in SalePrices.
4) Similary, with Abnorml Abnormal Sale - may be due to trade,short sale, foreclosure

Problem 4 : Any change in SalePrice over the period from 2006 to 2010 based on GarageQual

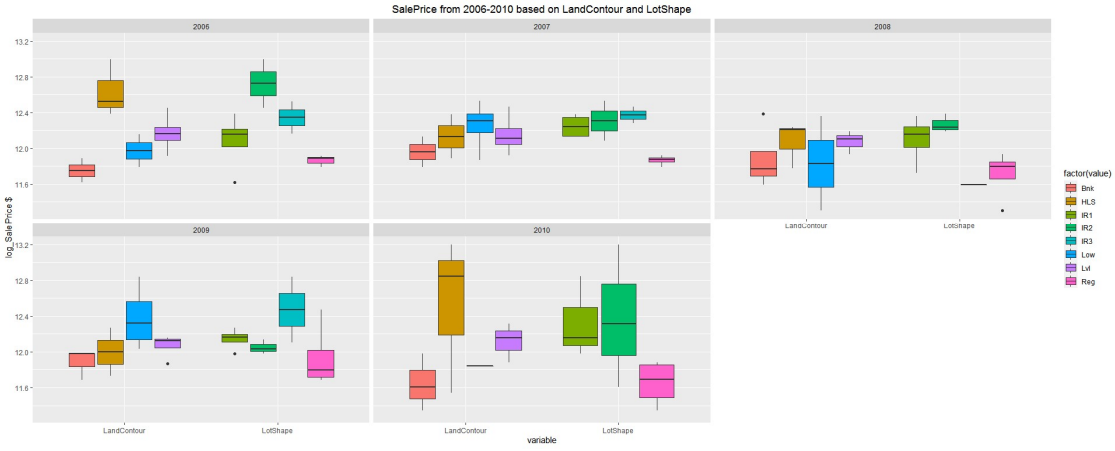


1) Definitely there was a change of SalePrice over years. During 2006 to 2009 2) The log_SalePrice for Average qualaty Garage properties sold around for 3500$ 3) In 2010 suddenly there is drop in SalePrice for even the Garage with excellent quality, may be due to the age of the Garage.

Problem 5:Over the years.How the SalePrice changed based on Neighborhood and BldgType .Explain

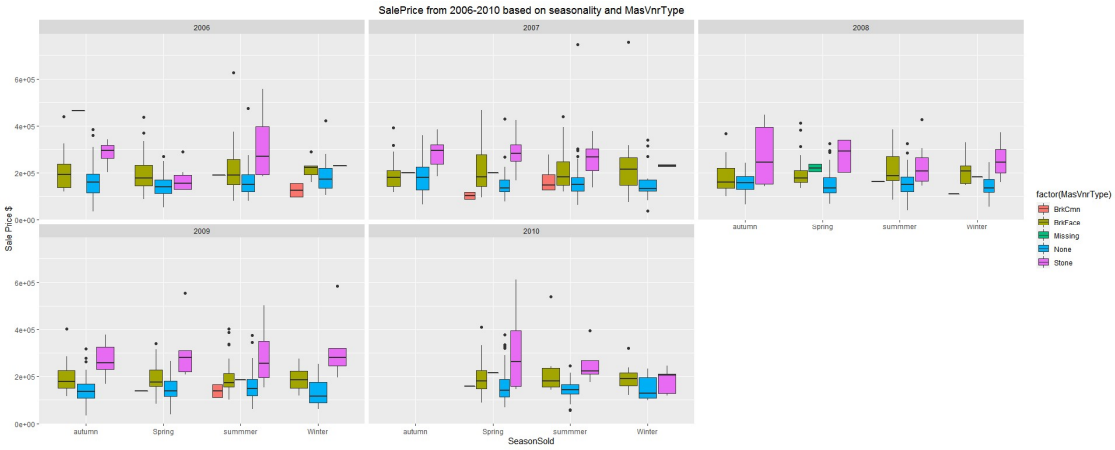
1) Buiding Type 1Fam has been sold in Neighborhoods are BrkSide, ClearCr, Gilbert, IDOTRR, NoRidge, NWAmes,Timber, NRidgHt, StoneBr and so on 2) In ClearCr Gilbert,Timber NoRidge location only buildingtype of 1Fam looks like trendy throughout from 2006 to 2010 The Sale price is almost Same in every year. 3) In Suburb Blmngtn CollgCr StoneBr, Somerst the building type TwnhsE, 1Fam looks like trendy. The saleprice is almost same for both building type from 2006-2010

Problem 6:Was there any difference in SalePrice for the properties sold between 2006 to 2010 based on LotShape/LandContour?which is basically considered in Indian tradition, Just for curiosity have included this problem analysis.



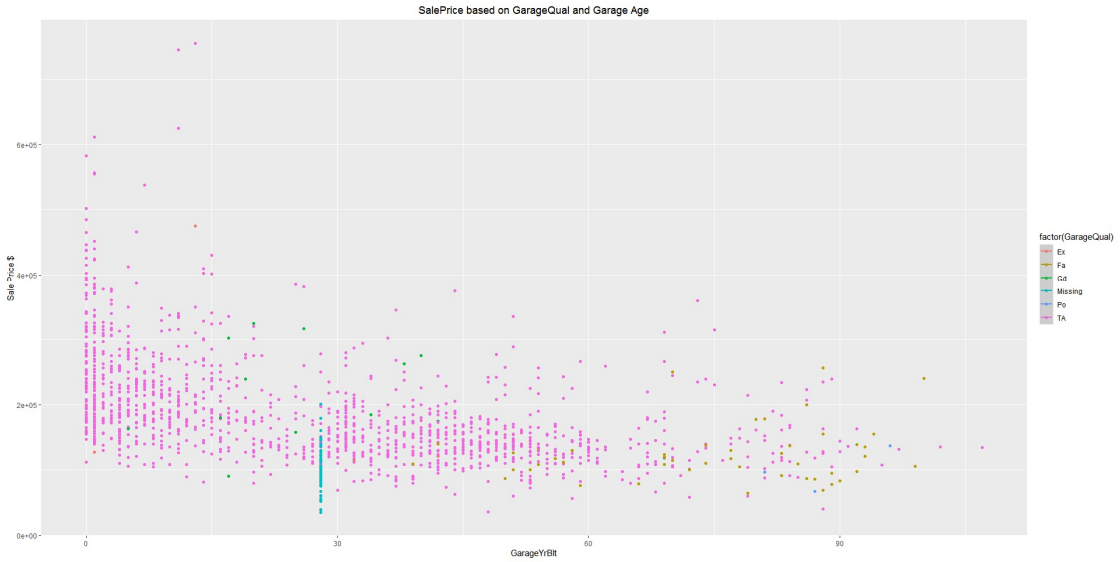
Looks like LandContour and Landshape also has prominent dependency on SalePrice in 2006, 2009 and 2010 Which are two of the majour features to be considered for modeling.

Problem 7:Check the seasonality of SalePrice based on MasVnrType used.



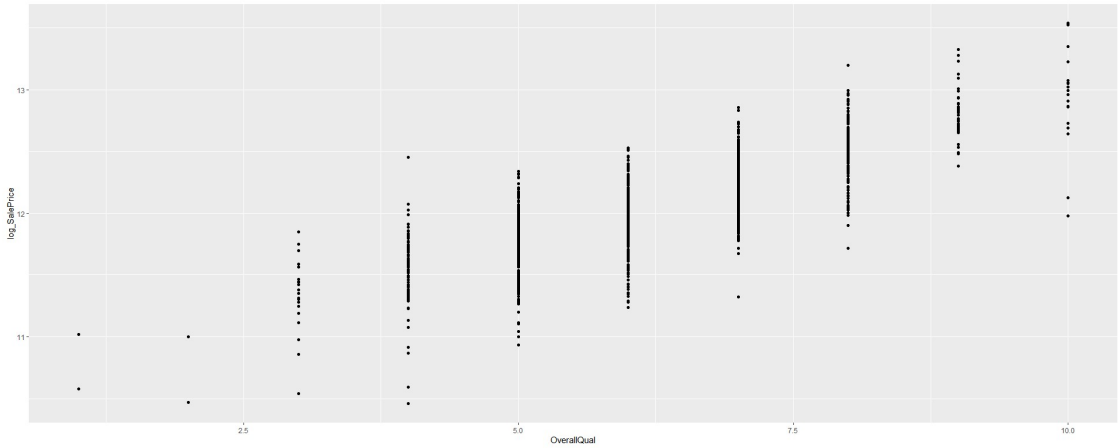
Very Slight Variation of SalePrice when it comes to MasVnrType Saleprice almost stays similar in all seasons with no much variation here.

Problem 8:Did the Age of Garage, property made any difference in the Saleprice from 2006 2010.



Definitely there is strong relation which negative. As the Garage Ages The SalePrice is decreasing. Most of the Garages also are typical/Average quality from the beginning of the 2006 data.

Problem 9:Do you think we could get good linear model with just considering Overall quality as a stand alone parameter for Sale Price prediction? Give thoughts




```
model <- lm(Ames_train_tidied$log_SalePrice ~ Ames_train_tidied$OverallQual,
            Ames_train_tidied)
summary(model)
```

Even though RMSE is 0.2303 the R-squared: 0.6678, which cannot be considered as a good linear model just by considering the OverallQuality.

The problem 10 will be solved once fitting the model is completed.

8. Linear Modeling:

Here considered features with strong correlation and good variability to SalePrice.

With the first fitted model, was able to achieve the R square=0.8599 and RMSE=0.1519 value.

```
summary(model_Ames_train_tidied)
```

With the second fitted model, was able to achieve the R square=0.8772 and RMSE=0.1431 value.

With the third fitted model, was able to achieve the R square=0.8507 and RMSE=0.1567 value.

```
  summary(mode3_Ames_train_tidied)
```

mode4_Ames_train_tidied is the best fit with RMSE=0.1401 # and R^2=0.884 compared to other fitted models for test data set-> lm

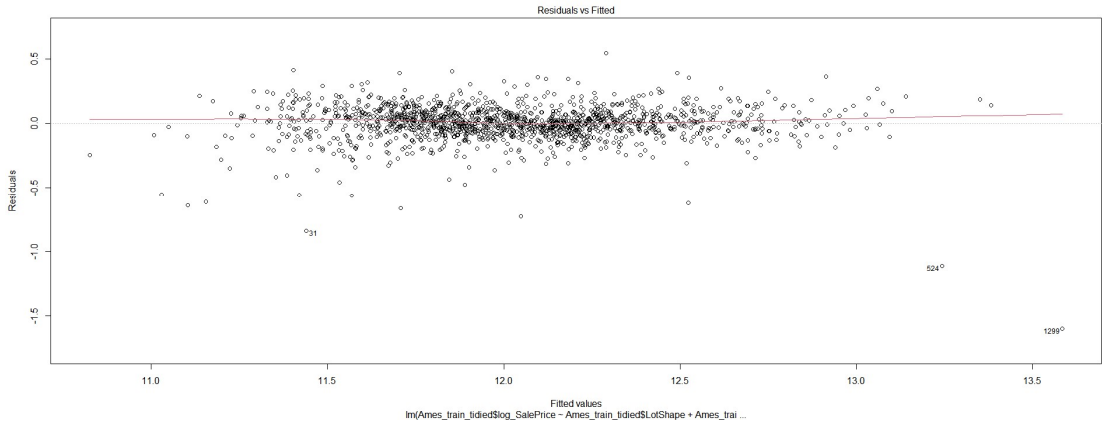
There are few outliers as following residual shows in the fourth model.

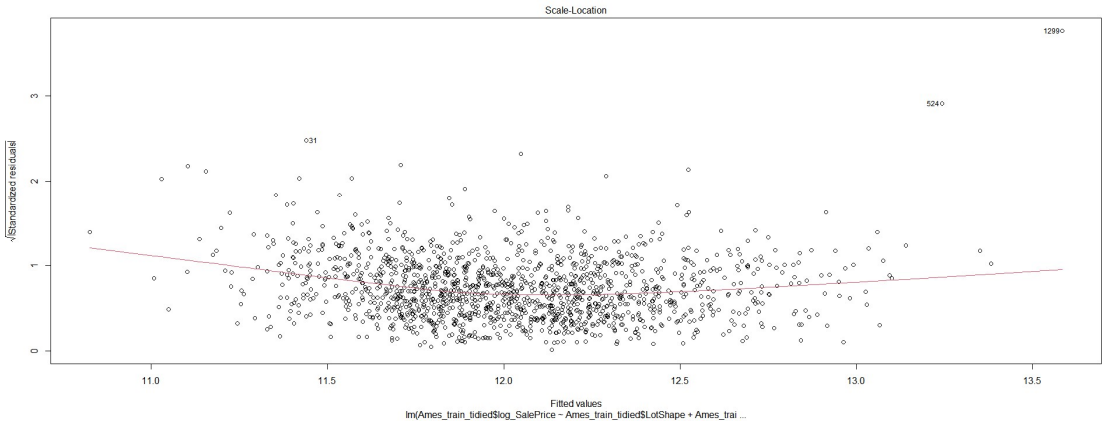
```
  outlierTest(mode4_Ames_train_tidied)

FALSE        rstudent unadjusted p-value Bonferroni p
FALSE 1299 -15.341654         3.7451e-49   5.4678e-46
FALSE 524   -8.731117         7.1822e-18   1.0486e-14
FALSE 31    -6.221352         6.5301e-10   9.5339e-07
FALSE 633   -5.451498         5.9142e-08   8.6348e-05
FALSE 463   -4.854481         1.3450e-06   1.9638e-03
FALSE 496   -4.780196         1.9394e-06   2.8316e-03
FALSE 1325  -4.596458         4.6921e-06   6.8505e-03
FALSE 969   -4.511834         6.9754e-06   1.0184e-02
FALSE 314    4.282135         1.9796e-05   2.8903e-02

  #There are few outliers.
```

With the third fitted model, was able to achieve the R square=0.884 and RMSE=0.1401 value.

```
  summary(mode4_Ames_train_tidied)
```

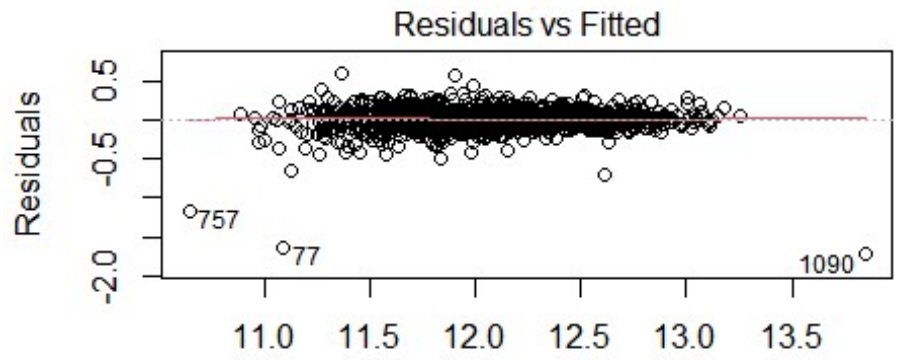
Here is the linear regression,residual and cook's distance plot for the best fitted model.



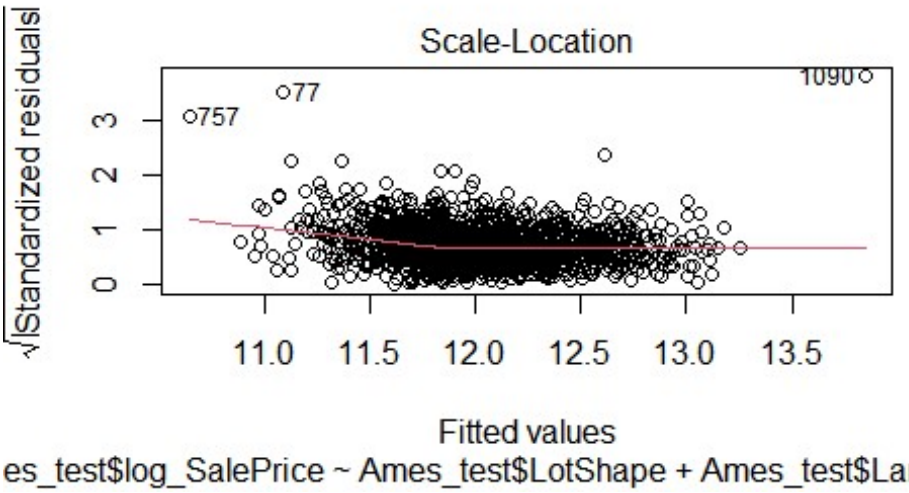


When predicted with test dataset, I got the R square-0.8973 and RMSE=0.1375 value which is very near to the fitted model.

```
summary(model_Ames_test)

plot(model_Ames_test,1)
```



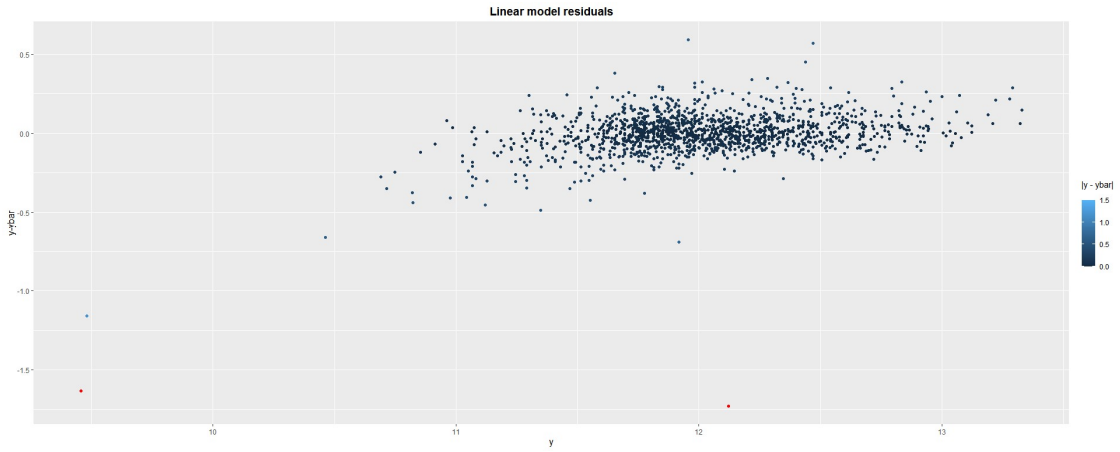

```
  plot(model_Ames_test,3)
```



Here is the comparision of RMSE values of both fitted model and test dataset.

```
  rmse(mode4_Ames_train_tidied,Ames_train_tidied)
FALSE [1] 0.1359736
  rmse(model_Ames_test,Ames_test)
FALSE [1] 0.1333493
```

predicting test dataset and plotting residuals



Once again checking here on the RMSE value of residuals.

```
FALSE [1] 0.1333493
```

Problem 10: Use predictions from your final model to compare suburbs which have shown varying growth. Or, to identify which suburbs have been growing the most over the last few years.

According the fitted final model the formula for prediction of SalePrice is as follows.

*SalePrice = (-3.479e-02)*NeighborhoodBlueste+1.318e+01*

*SalePrice = (-9.795e-02)*NeighborhoodBrDale+1.318e+01*

*SalePrice = (1.237e-01)*NeighborhoodVeenker+1.318e+01*

and so on

so Based on the Neighbourhood value in test dataset SalePrice will be predicted. Similary the formulas could be derived with each and every dependent feature to predict SalePrice. The final formula would be as follows.

as Multiple linear models will follow the general form

y = a1x1 + a2x2 + . + b

SalePrice = a1*Neighborhood + a2*OverallQual + b….and so on

while a1= feature1*coefficient+/-intercept and so on wheere feature=Neighborhood and coefficient and intercepts are the respective estimations using linear modeling.

Conclusion:

By running analysis gathering and including multiple methods of data processing and alalysis techniques, I have determined an acceptable multiple linear regression model to

the dataset. Firstly, I created 4 models with train dataset which I tidied. Based on the characteristics of the features I could create 4 models. ##### The first model having R square value of 0.8599 and RMSE: 0.1519. ##### The second model having R square value of 0.8772 and RMSE: 0.1431. ##### The third model having R square value of 0.8507 and RMSE: 0.1567. ##### The fourth model having R square value of 0.884 and RMSE: 0.1401 I basically selected features based on strong correlation with the SalePrice and Variabilty in the distribution to the SalePrice. Finally, I choose fourth model as the best model to fit and predit the values. Later, predicted with test dataset with the best model and plotted residuals, to check whether the prediction based on the fitted model is near to the absolute one. Here considered log_SalePrice instead of SalePrice. With SalePrice considered for modeling gave RMSE=30020 and R-squared=0.8654, it is best to use log transformation of SalePrice.

If any future work conducted, I would like to work on multiple regression techniques, not only on linear regression. That would help me analyze and consider the rest of the features which I dropped here while doing linear regression decisions.

References: Big vote of thanks to all the references mentioned below here. Without which I would have not successfully able to complete linear modelling for multivariable data set.

http://jse.amstat.org/v19n3/decock.pdf http://stackoverflow.com/
https://rpubs.com/RobbyS/622233
https://scholarworks.calstate.edu/downloads/fx719m836
https://www.youtube.com/watch?v=wR4Xfwjr-3Y&list=LL&index=14
https://nycdatascience.com/

Note: As there was limited number of pages to be submitted, few of the plots are not printed. Please enable them to check or check the R code submitted along with the RMarkdown file. As the report file was with 25 pages, I tried to reduce printing the number of plots and font size.